

STATISTICAL INFERENCE BASED ON ROBUST LOW-RANK DATA MATRIX APPROXIMATION

BY XINGDONG FENG¹ AND XUMING HE²

Shanghai University of Finance and Economics and University of Michigan

The singular value decomposition is widely used to approximate data matrices with lower rank matrices. Feng and He [*Ann. Appl. Stat.* **3** (2009) 1634–1654] developed tests on dimensionality of the mean structure of a data matrix based on the singular value decomposition. However, the first singular values and vectors can be driven by a small number of outlying measurements. In this paper, we consider a robust alternative that moderates the effect of outliers in low-rank approximations. Under the assumption of random row effects, we provide the asymptotic representations of the robust low-rank approximation. These representations may be used in testing the adequacy of a low-rank approximation. We use oligonucleotide gene microarray data to demonstrate how robust singular value decomposition compares with the its traditional counterparts. Examples show that the robust methods often lead to a more meaningful assessment of the dimensionality of gene intensity data matrices.

1. Introduction. Research on robustness dates back to the prehistory of statistics. However, the concepts and theories of robust statistics have not been formally and systematically established until recent decades [Huber and Ronchetti (2009), Hampel et al. (1986)]. Much work on robust statistics has focused on linear regression and multivariate location-scatter models. It has been well recognized that the least squares method under those models is sensitive to a small number of outliers. Robust methods are generally developed to down-weight outliers.

Received August 2013.

¹Supported by NNSF of China Grant 11101254, Shanghai Pujiang Program, Program for Changjiang Scholars and Innovative Research Team in Universities, and Key Laboratory of Mathematical Economics (SUFU), Ministry of Education, China.

²Supported by NSF Grants DMS-13-07566, DMS-12-37234, NIH Grant R01GM080503 and NNSF of China Grant 11129101.

AMS 2000 subject classifications. Primary 62F03, 62F35; secondary 62F05, 62F10, 62F12.

Key words and phrases. Hypothesis testing, M estimator, singular value decomposition, trimmed least squares.

This is an electronic reprint of the original article published by the [Institute of Mathematical Statistics](#) in *The Annals of Statistics*, 2014, Vol. 42, No. 1, 190–210. This reprint differs from the original in pagination and typographic detail.

The singular value decomposition (SVD) of a data matrix is often used as a data reduction tool. In fact, the SVD can be viewed as a basic tool in dimension reduction. Consider a data matrix

$$\mathbf{Y} = \begin{pmatrix} y_{11} & \cdots & y_{1m} \\ \vdots & \vdots & \vdots \\ y_{n1} & \cdots & y_{nm} \end{pmatrix},$$

of n rows and m columns, where m is fixed. An approximation of rank r to the matrix can be found by

$$(1) \quad \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - z_{ij})^2,$$

where z_{ij} are the elements of $\mathbf{Z} = \mathbf{R}\mathbf{C}$ for an $n \times r$ matrix \mathbf{R} and a $r \times m$ matrix \mathbf{C} . The matrices \mathbf{R} and \mathbf{C} are not identifiable in this formulation, so additional constraints may be imposed to ensure identifiability. As pointed out in Ammann (1993) and Chen, He and Wei (2008), the SVD is equivalent to the least squares approach to a bilinear regression model, so it suffers from the usual lack of robustness against outliers.

Ruppert and Carroll (1980) have used the trimmed least squares estimation in the linear model by using weights obtained from some initial consistent estimates, and Gervini and Yohai (2002) have considered a variant of the trimmed method leading to the maximum breakdown point and full asymptotic efficiency under normal errors. In this paper, we adopt the idea of using trimmed least squares estimation, where the scheme of choosing weights is explained in Section 2. The low-rank approximation of matrices by weighted least squares has been considered by Gabriel and Zamir (1979), but their weights are fixed, while the weights of the proposed method in this paper are obtained from an initial robust estimate.

We will consider a two-step approximation method in this paper. More specifically, we consider the first approximation by minimizing

$$(2) \quad \sum_{i=1}^n \hat{w}_i \sum_{j=1}^m \left(y_{ij} - \sum_{k=1}^r \theta_{ki} \phi_{kj} \right)^2,$$

where \hat{w}_i are the weights based on an initial estimate (to be described later), θ_{ki} are the elements of \mathbf{R} , and ϕ_{kj} are the elements of \mathbf{C} . However, it is clear that the estimates of θ 's are the linear combination of vectors \underline{y} 's given ϕ 's, so it implies that this lower-rank approximation is not robust against outliers. Then we consider the second approximation by using the estimated ϕ 's from the first step, denoted collectively as $\tilde{\phi}_k$ ($k = 1, \dots, r$), and then minimizing

$$(3) \quad \sum_{i=1}^n \sum_{j=1}^m L \left(y_{ij} - \sum_{k=1}^r \theta_{ki} \tilde{\phi}_{kj} \right),$$

over the θ 's for some robust loss function L , where $\tilde{\phi}_{kj}$ is the j th component of $\tilde{\phi}_k$. Our statistical analysis will be performed under the following model:

$$(4) \quad \underline{y}_i = \sum_{k=1}^r \theta_{ki}^{(0)} \underline{\phi}_k^{(0)} + \varepsilon_i, \quad i = 1, \dots, n,$$

where $\underline{y}_i = (y_{i1}, \dots, y_{im})^T$ is the i th observed vector, $\underline{\theta}_k^{(0)} = (\theta_{k1}^{(0)}, \dots, \theta_{kn}^{(0)})^T$ is used to explain the row effects, and $\underline{\phi}_k^{(0)} = (\phi_{k1}^{(0)}, \dots, \phi_{km}^{(0)})^T$ is used to explain the column effects in the data matrix. The row effects $\theta^{(0)}$'s are assumed to be random, and the length of observed vectors, m , is fixed. We are interested in the structure of the mean matrix $E(\mathbf{Y})$, and the uniqueness of the low-rank representation is implied by conditions (M1) and (M2) given in Appendix A.1. In our work, we assume that each component of ε_i in the model is symmetrically distributed, but outliers might be present in the data. The robust methods are meant to be reliable against violations of the model assumptions.

Our model includes that of Feng and He (2009) as a special case where ε_i is Gaussian. The main contribution of the present paper is to develop a robust procedure that can accommodate outlying measurements in the data matrix. To achieve this goal, we have to utilize nonlinear operations in the estimation procedure, and consequently, we need to analyze the statistical properties of the robust procedures with a new set of techniques.

When the data matrix is the sum of low-rank and sparse matrices, the theory of the exact recovery of both matrices has been established by Candès et al. (2011) and Zhou et al. (2009). Agarwal, Negahban and Wainwright (2012) further consider a broader class of models, where random errors are introduced, and the penalized method is used for estimation. These authors have provided deterministic error bounds for their estimates while allowing the number of columns to grow with n , but in this paper we are interested in hypothesis testing based on the asymptotic representation of the robust estimates with a fixed number of columns.

For the estimates of $\underline{\theta}_k^{(0)}$ and $\underline{\phi}_k^{(0)}$ ($k = 1, \dots, r$) obtained from (3) and (2), respectively, we shall derive their asymptotic representations in Section 3 as $n \rightarrow \infty$. In Appendix B, we discuss some finite sample properties of the estimators (3), which are critical for the theoretical development in Section 4, where we robustify the tests of unidimensionality for testing the adequacy of a unidimensional model against the alternative rank-two mean structure for the data matrix. In Section 4.3, we compare the results of testing unidimensionality of matrices from Feng and He (2009) with those from the robust alternative in microarray data analysis. Technical assumptions of our model are given in Appendix A, and the proofs for the lemmas and the theorems given in the paper can be found in Appendix B and in the supplementary material [Feng and He (2014)].

2. Estimation procedure. In this paper, we propose the following procedure to estimate the row and column parameters of model (4).

Step 0. Construction of an initial robust estimate of column parameters:

- (I1) prechoose a constant α^* (typically between 0.1 and 0.5);
- (I2) select $\lceil (1 - \alpha^*)n \rceil$ rows randomly from the data matrix, and denote this matrix as \mathbf{Y}^* , where $\lceil x \rceil$ is the smallest integer greater than x ;
- (I3) carry out the regular SVD on the matrix \mathbf{Y}^* and obtain the first r right singular vector as $\hat{\phi}_k$, $k = 1, \dots, r$;
- (I4) estimate the row parameters $\theta_k^{(0)}$ by minimizing the objective function (3), in which $\tilde{\phi}_k$ is replaced by $\hat{\phi}_k$. The resulting estimate is denoted as $\hat{\theta}_k$;
- (I5) repeat (I2)–(I4) for a prespecified number of times (to be discussed later), and find the subset of $\lceil (1 - \alpha^*)n \rceil$ rows that gives the minimum value of $\sum_{i=1}^n \sum_{j=1}^m L(y_{ij} - \sum_{k=1}^r \hat{\theta}_{ki} \hat{\phi}_{kj})$.

Step 1. Computation of the weighted least squares to improve efficiency of the column parameters:

- (1a) given the initial estimate of the column parameters, choose a trimming proportion $\alpha \leq \alpha^*$ and calculate the weights

$$(5) \quad \hat{w}_i = 1(\hat{\xi}_\alpha < \|\hat{\underline{e}}_i\|^2 \leq \hat{\xi}_{1-\alpha}),$$

where $\hat{\xi}_\alpha$ is the sample α quantile of $\|\hat{\underline{e}}_i\|^2$ and $\hat{\underline{e}}_i = (I_m - \sum_{k=1}^r \hat{\phi}_k \hat{\phi}_k^T) \underline{y}_i$;

- (1b) given the weights, obtain the estimate $\tilde{\phi}_k$ ($k = 1, \dots, r$) of the column parameters by minimizing (2) over the row and column parameters.

Step 2. Updating row effect estimates with robustness: given $\tilde{\phi}_k$ from step 1, obtain the estimate $\tilde{\theta}_k$ of the row effects by minimizing (3) over the row parameters.

In step 0, we obtain an initial root- n robust estimate of the column parameters $\phi_k^{(0)}$, denoted as $\hat{\phi}_k$, $k = 1, \dots, r$. The choice of α^* should reflect what percentage of outlying rows we expect, and it is similar to the amount of trimming one chooses to use in the trimmed mean. The number of subsets used in (I5) is fixed and should be chosen to ensure that there is a high probability that one of the subsets contains no outliers. For example, if we have 20 rows in the data matrix and expect 2 outlying rows, by choosing $\alpha^* = 0.3$ to use subsets of 14 rows, the probability that one random subset is outlier-free is nearly 0.08. If we use 100 random subsets in (I5), the probability of having at least one outlier-free subset is greater than 0.999. Simple calculations like this show that we can obtain a robust estimate through this procedure with high probability.

Because the estimate $\hat{\underline{\phi}}_k$ in (I3) is the least squares estimate considered in Feng and He (2009), and the size of the subset is proportional to n , then the initial estimate of column vectors here is root- n consistent. Given the initial estimate of the column parameters, we calculate the weights in step (1a), where the trimming level α plays the same role as α^* in (I1) but in a different context. The main purpose of step 1 is to increase efficiency of the column parameter estimates over those from step 0, but the corresponding estimates of the row effects might not be robust. The purpose of step 2 is to robustify the row effect estimates.

General weight functions of $\|\hat{\underline{e}}_i\|^2$ can be considered in lieu of (5), but we expect that the results given in the Appendix B still hold under appropriate regularity conditions. Our proposed robust estimates of parameters $\underline{\theta}_k^{(0)}$ and $\underline{\phi}_k^{(0)}$ are obtained by minimizing (3) and (2), respectively. By considering the regular SVD on the approximation matrix $\sum_{k=1}^r \tilde{\underline{\theta}}_k \tilde{\underline{\phi}}_k^T$, we actually obtain a robust SVD on the data matrix \mathbf{Y} .

3. Asymptotic properties. The data matrix \mathbf{Y} often arises with the rows representing individuals randomly sampled from a large population, but the columns for measurements at m different locations or time points. It is then natural to use $\underline{\theta}$ as the random row effects, and $\underline{\phi}$ as the fixed column effects. Individuals can be characterized by the row effects, and their spatial or temporal profiles can be understood by the column effects. The distinction between the random and the fixed effects is not relevant to the optimization problems (2) and (3) themselves, but is important for the statistical properties of the estimates obtained from the optimization. To derive the statistical representations of the row- and column-effect estimates, we use conditions (M1)–(M5) detailed in Appendix A.1. Those conditions also ensure proper parameter identifications.

Following Definition 1.1 of Feng and He (2009), we use the rank of the mean matrix $E(\mathbf{Y})$ as the dimensionality of the model. A unidimensional model refers to the mean matrix of rank-one. For unidimensional data, we can use the first singular component to summarize the row and column effects. For example, if a unidimensional test of m items is given to n examinees, the data matrix as the scores of the examinees on each of the items might be expected to be of rank one, where a rank-one approximation uses θ_i to summarize the “ability” of the i th examinee and ϕ_j to represent the difficulty level of the j th item. In educational measurements, different forms of unidimensionality has been used. For a related article on assessing unidimensionality of polytomous data, see Nandakumar et al. (1998).

3.1. Profiling in optimization and column effect estimates. The number of the θ ’s involved in the objective function (2) increases with n , which inconveniences the asymptotic analysis as $n \rightarrow \infty$. To bypass this difficulty,

we view $\underline{\theta}_k$ as nuisance parameters in the following profiling procedure. First, we minimize the objective function (2) with respect to $\underline{\theta}_k$ as if $\underline{\phi}_k$ ($k = 1, \dots, r$) were known. Then, with the estimates $\theta_{ki}^* = \underline{\phi}_k^T \underline{y}_i$, minimizing (2) is equivalent to minimizing the following objective function:

$$(6) \quad \min_{\underline{\phi}} \sum_{i=1}^n \hat{w}_i \left\| \left(I_m - \sum_{k=1}^r \underline{\phi}_k \underline{\phi}_k^T \right) \underline{y}_i \right\|^2,$$

under the restrictions that $\|\underline{\phi}_1\| = \dots = \|\underline{\phi}_r\| = 1$, and $\underline{\phi}_k \perp \underline{\phi}_l$ for $k \neq l$.

Let $\varphi_0 = (\underline{\phi}_1^{(0)T}, \dots, \underline{\phi}_r^{(0)T})^T$, $\vartheta_0 = (0, \varphi_0^T)^T$, and $\hat{\vartheta}_\tau = (\hat{\xi}_\tau - \xi_\tau, \hat{\varphi}^T)^T$, where ξ_τ is the τ th quantile of $\|\underline{e}_i\|^2$, $\underline{e}_i = (I_m - \sum_{k=1}^r \underline{\phi}_k^{(0)} \underline{\phi}_k^{(0)T}) \underline{y}_i$ and $\hat{\varphi}$ is the initial estimate of φ_0 . We obtain the Bahadur representation for the estimates $\tilde{\varphi} = (\tilde{\phi}_1^T, \dots, \tilde{\phi}_r^T)^T$ from step 1.

THEOREM 3.1. *Assume model (4) with $\hat{\varphi}$ as any root- n consistent estimate of the parameter vector φ_0 . If conditions (M1)–(M5) and (E1)–(E3) in Appendix A hold, then*

$$(7) \quad \begin{aligned} \tilde{\varphi} - \varphi_0 = & -(n\mathbf{D}_0)^{-1} \sum_{i=1}^n w_i \begin{pmatrix} \underline{b}_1(\theta_{1i}^{(0)}, \dots, \theta_{ri}^{(0)}, \underline{\varepsilon}_i, \varphi_0) \\ \vdots \\ \underline{b}_r(\theta_{1i}^{(0)}, \dots, \theta_{ri}^{(0)}, \underline{\varepsilon}_i, \varphi_0) \end{pmatrix} \\ & + \mathbf{G}_n^T \begin{pmatrix} \hat{\vartheta}_{1-\alpha} - \vartheta_0 \\ \hat{\vartheta}_\alpha - \vartheta_0 \end{pmatrix} + o_p(n^{-1/2}), \end{aligned}$$

where $w_i = 1(\xi_\alpha < \|\underline{e}_i\|^2 \leq \xi_{1-\alpha})$, \mathbf{D}_0 is an $mr \times mr$ nonsingular square matrix, \mathbf{G}_n is an $mr \times 2(mr+1)$ matrix with the Frobenius norm $\|\mathbf{G}_n\|_F = O(1)$ and

$$\begin{aligned} & \underline{b}_j(\theta_{1i}^{(0)}, \dots, \theta_{ri}^{(0)}, \underline{\varepsilon}_i, \varphi_0) \\ &= 2\{\theta_{ji}^{(0)} + \underline{\varepsilon}_i^T \underline{\phi}_j^{(0)}\}^2 \underline{\phi}_j^{(0)} - \{\theta_{ji}^{(0)} + \underline{\varepsilon}_i^T \underline{\phi}_j^{(0)}\} \underline{y}_i - \sum_{k=1}^r \{\theta_{ki}^{(0)} + \underline{\varepsilon}_i^T \underline{\phi}_k^{(0)}\} \underline{y}_i. \end{aligned}$$

The specific forms of \mathbf{D}_0 and \mathbf{G}_n can be found in the supplementary material [Feng and He (2014)]. From Theorem 3.1, Lemma B.2 (in the Appendix) and Theorem 2.2 of Feng and He (2009), it is clear that the estimate $\tilde{\varphi}$ of the parameter vector φ_0 is root- n consistent with asymptotic normality. Its asymptotic variance-covariance matrix is complicated because both variations from the initial estimates and the variation from the weighted least squares method are present.

3.2. Row effect predictions. Note that the least squares estimate of $\theta_{ki}^{(0)}$ is $\tilde{\phi}_k^T \underline{y}_i$, so it can be seriously affected by any outlying value of the observed

vector \underline{y}_i . We now consider the robust procedure that minimizes (3) for a smooth loss function L .

If L has continuous second derivative, the minimizers of (3) are, by the implicit function theorem in calculus,

$$\begin{aligned} (8) \quad \tilde{\theta}_{1i} &= f(\underline{y}_i, \tilde{\phi}_1, \tilde{\phi}_2, \dots, \tilde{\phi}_r), \\ &\vdots \\ (9) \quad \tilde{\theta}_{ri} &= f(\underline{y}_i, \tilde{\phi}_r, \tilde{\phi}_1, \dots, \tilde{\phi}_{r-1}), \end{aligned}$$

where f is a function with continuous partial derivatives with respect to ϕ_{kj} for $k = 1, \dots, r$ and $j = 1, \dots, m$.

Before we move on, it helps to explore some properties of the implicit function f . Consider minimizing the following objective function:

$$\sum_{j=1}^m L \left(y_{ij} - \sum_{k=1}^r \theta_{ki} \phi_{kj}^{(0)} \right),$$

which can be written under model (4) as

$$\sum_{j=1}^m L \left(\varepsilon_{ij} - \sum_{k=1}^r (\theta_{ki} - \theta_{ki}^{(0)}) \phi_{kj}^{(0)} \right).$$

When this minimization is performed with respect to θ_{ki} , we have

$$(10) \quad f(\underline{y}_i, \phi_1^{(0)}, \phi_2^{(0)}, \dots, \phi_r^{(0)}) = \theta_{1i}^{(0)} + f(\underline{\varepsilon}_i, \phi_1^{(0)}, \phi_2^{(0)}, \dots, \phi_r^{(0)}),$$

$$\vdots$$

$$(11) \quad f(\underline{y}_i, \phi_r^{(0)}, \phi_1^{(0)}, \dots, \phi_{r-1}^{(0)}) = \theta_{ri}^{(0)} + f(\underline{\varepsilon}_i, \phi_r^{(0)}, \phi_1^{(0)}, \dots, \phi_{r-1}^{(0)}).$$

If L is even, then the function f is radially symmetrical with respect to its first argument. We obtain the asymptotic result for the estimates θ_{ki} defined as the minimizer of (3) in the following theorem.

THEOREM 3.2. *Assume model (4) with $\hat{\varphi}$ as any root- n consistent estimate of the parameter vector φ_0 . If conditions (M1)–(M5), (A1)–(A4) and (C3) in Appendix A hold, then*

$$\begin{aligned} \sum_{k=1}^r \tilde{\theta}_{ki} \tilde{\phi}_k &\xrightarrow{d} \sum_{k=1}^r \theta_{ki}^{(0)} \phi_k^{(0)} \\ &\quad + f(\underline{\varepsilon}_i, \phi_1^{(0)}, \phi_2^{(0)}, \dots, \phi_r^{(0)}) \phi_1^{(0)} \\ &\quad + \dots + f(\underline{\varepsilon}_i, \phi_r^{(0)}, \phi_1^{(0)}, \dots, \phi_{r-1}^{(0)}) \phi_r^{(0)}, \end{aligned}$$

where $\tilde{\theta}_{ki}$ is defined in (8)–(9), and \xrightarrow{d} refers to convergence in distribution.

It is clear from Theorem 3.2 that each row of the approximating matrix $\sum_{k=1}^r \tilde{\theta}_k \tilde{\phi}_k^T$ converges in distribution to the corresponding row of the rank- r matrix $\sum_{k=1}^r \theta_k^{(0)} \phi_k^{(0)T}$ and some function of the model errors $\underline{\varepsilon}$.

4. Application. For vector measurements, a unidimensional summary is widely used in data analysis. In this section, we consider testing on the sufficiency of unidimensional summaries, against the alternative that the matrix \mathbf{Y} is a rank two matrix under model (4).

4.1. Hypothesis testing. With the asymptotic results of the previous section, we consider hypothesis testing here based on the robust estimates. The null hypothesis is $\underline{\mu}_2 = \underline{0}$, which implies unidimensionality of the mean matrix $E(\mathbf{Y})$, and that no meaningful pattern can be found in the second dimension of the data matrix. This hypothesis is especially interesting in the probe-level microarray data analysis, where unidimensional models are usually assumed to summarize the gene expression level from the intensity data matrix [Li and Wong (2001), Irizarry et al. (2003)].

We first consider the estimation by minimizing (2) with $r = 2$. We then use the column vectors $\tilde{\phi}_1$ and $\underline{0}$ in minimizing (3) to obtain the estimate $\tilde{\theta}_{1i} = f(\underline{y}_i, \tilde{\phi}_1, \underline{0})$, where f is defined in (8)–(9). For convenience, we use $f(\underline{y}_i, \tilde{\phi}_1)$ instead of $f(\underline{y}_i, \tilde{\phi}_1, \underline{0})$ from now on. Let

$$\gamma(\underline{y}_i, \varphi) = \sum_{j=1}^m L'(y_{ij} - f(\underline{y}_i, \phi_1)) \phi_{1j} \phi_{2j}$$

be the score for unidimensionality corresponding to the i th vector \underline{y}_i . We have the following result.

THEOREM 4.1. *Let $\underline{a} = (a_1, \dots, a_n)^T$ be a vector that is orthogonal to $\underline{\mu}_1$ and satisfies $\|\underline{a}\|^2 = n$ with a bounded supremum norm. Assume model (4) and conditions (M1)–(M5), (C1)–(C4), (D1)–(D2) in Appendix A, then*

$$(12) \quad n^{-1/2} \underline{a}^T \tilde{\gamma} / \tilde{\sigma}_n \xrightarrow{L} N(0, 1),$$

under the null hypothesis that $\underline{\mu}_2 = \underline{0}$, where

$$(13) \quad \tilde{\gamma} = (\gamma(\underline{y}_1, \tilde{\varphi}), \dots, \gamma(\underline{y}_n, \tilde{\varphi}))^T,$$

$$(14) \quad \tilde{\sigma}_n^2 = n^{-1} \sum_{i=1}^n \gamma^2(\underline{y}_i, \tilde{\varphi}) - \left\{ n^{-1} \sum_{i=1}^n \gamma(\underline{y}_i, \tilde{\varphi}) \right\}^2$$

and $\tilde{\varphi}$ is the robust estimate defined in Section 2.

REMARK 4.1. If the loss function L is the L_2 norm, then $L'(x) = 2x$. It then follows that $\gamma(\underline{y}_i, \tilde{\varphi}) = 2 \sum_{j=1}^m \{y_{ij} - (\tilde{\phi}_1^T \underline{y}_i) \tilde{\phi}_{1j}\} \tilde{\phi}_{2j} = 2 \tilde{\phi}_2^T \underline{y}_i$, because

$\tilde{\phi}_1 \perp \tilde{\phi}_2$ for the least squares case. Thus, the statistic used by Feng and He (2009) can be viewed as a special case of Theorem 4.1.

If the direction vector \underline{a} is not orthogonal to $\underline{\mu}_1$, then $n^{-1/2}\underline{a}^T\tilde{\gamma}/\tilde{\sigma}_n$ may not converge in distribution to a mean zero distribution. Typically $\underline{\mu}_1$ is unknown and needs to be estimated. This is usually done by extra group information in the rows to enable us to consistently estimate $\underline{\mu}_1$, which is sufficient to have the asymptotic result for the pivotal statistic in Theorem 4.1. This theorem also ensures the validity of the bootstrap as described in Section 3.3 of Feng and He (2009) based on Theorem 1 of Mammen (1991).

It is certainly possible that the direction vector \underline{a} happens to be a poor choice in the sense of low power against a particular alternative. To ensure decent power of the test, we can consider several target directions that are orthogonal to each other.

THEOREM 4.2. *Assume the conditions of Theorem 4.1. Consider a $K \times n$ matrix A with all the row vectors orthogonal to each other, with K being fixed. If the vector $\underline{a}_l = (a_{l1}, \dots, a_{ln})^T$ is the l th row of the matrix A and satisfies $\underline{a}_l \perp \underline{\mu}_1$, and $\|\underline{a}_l\|^2 = n$ with uniformly bounded elements, then $P(n^{-1}\|A\tilde{\gamma}\|^2/\tilde{\sigma}^2 \leq x) - F_K(x) \rightarrow 0$ under the null hypothesis that $\underline{\mu}_2 = 0$, where F_K is the cumulative distribution function of the χ_K^2 distribution, $\tilde{\gamma}$ and $\tilde{\sigma}$ are given in (13) and (14), respectively.*

4.2. A simulation study. In this section, we use a simulation study to assess the performance of the target direction test based on robust loss functions. We independently generate 20 rows of size 12 from model (4), with the mean of the corresponding 20×12 matrix equal to $\underline{\mu}_1\phi_1^T$ and $\underline{\mu}_1\phi_1^T + \underline{\mu}_2\phi_2^T$ under the null and the alternative hypotheses, respectively, where $\underline{\mu}_1 = (20, \dots, 20)^T$, $\underline{\mu}_2 = 2^{1/2}(1, -1, \dots, 1, -1)^T$, $\phi_1 = (1, \dots, 1)^T/12^{1/2}$ and $\phi_2 = (1, -1, \dots, 1, -1)^T/12^{1/2}$. The random effects $\theta_{1i}^{(0)} - \mu_{1i}$ and $\theta_{2i}^{(0)} - \mu_{2i}$ are generated from normal distributions with mean 0 and variances 4 and 1, respectively.

To assess the robustness of the method, we generate model errors in two ways. In an outlier-free model, all the errors are independently generated from one of the three cases: (I) $2^{-1/2}N(0, 1)$; (II) $(3/10)^{-1/2}t_5$, where t_5 is the t distribution with 5 degrees of freedom; (III) $2^{-1}(\chi_1^2 - 1)$, where χ_1^2 is the χ^2 distribution with 1 degree of freedom. In a contaminated model, the first two rows of the matrix are generated from the mixture of the normal distribution $N(0, 11)$ with probability 0.1 and one of the three distributions (I), (II) or (III) with probability 0.9, but the other rows are generated as in the outlier-free model. Under the contaminated model, outliers are likely to occur in the first two rows. A total of 5000 data sets are generated from each model in the simulation study.

TABLE 1

Estimated type I errors and powers of various tests at the nominal level of 5%, with data generated from outlier-free models

Size	Null			Alternative		
	Normal	t	χ^2	Normal	t	χ^2
Logistic ^a	0.051	0.049	0.043	1.000	0.999	0.995
Huber ^a	0.051	0.049	0.043	1.000	0.999	0.997
Least squares ^a	0.050	0.045	0.035	1.000	0.999	0.998
Logistic ^b	0.052	0.054	0.051	0.941	0.959	0.978
Huber ^b	0.053	0.054	0.051	0.936	0.956	0.976
Least squares ^b	0.054	0.046	0.039	1.000	0.989	0.998

^aThe results are from the case where $\underline{a} \propto \underline{\mu}_2$.

^bThe results are from the case where $\underline{a} \propto (3/2)^{1/2}(1, -1, \dots, 1, -1)^T + (1, \dots, 1, -1, \dots, -1)^T$.

For the initial steps (I1)–(I5) of Section 2, we use $\alpha^* = 0.3$ and 100 randomly selected subsets, and the constant $\alpha = 0.1$ is used in calculating the weights (5). With only two possible outlying rows, the probability that all 100 subsets contain an outlier is less than 0.001.

We consider two choices of the direction vector \underline{a} , with $\underline{a} \propto \underline{\mu}_2$ in the first case, and $\underline{a} \propto (3/2)^{1/2}(1, -1, \dots, 1, -1)^T + (1, \dots, 1, -1, \dots, -1)^T$ in the second case. The bootstrap calibration method of Feng and He (2009) is used to calculate the p -values of the tests. Three loss functions are used for comparison. They are

(L1) “Logistic”: $L(s) = C \log(\cosh(s/C))$,

(L2) “Huber”:

$$L(s) = \begin{cases} 2^{-1}s^2, & |s| \leq C, \\ C|s| - 2^{-1}C^2, & |s| > C, \end{cases}$$

(L3) “Least squares”: $L(s) = s^2$,

where $C = 0.1$ is used in our simulation. Since C is close to zero, the two robust loss functions (L1) and (L2) lead to results that are similar to those obtained under the L_1 loss $L(s) = |s|$.

We summarize the results for the outlier-free models in Table 1. It is clear from Table 1 that all the three tests preserve type I errors well, and they achieve very high power under the alternative. The story is different, however, for the contaminated models with the results in Table 2. When no more than 10% of outliers are present, the test based on the square loss becomes too conservative with low power, but the robust tests with (L1) and (L2) loss functions withstand the outliers very well.

TABLE 2
Estimated type I errors and powers of various tests at the nominal level of 5%, with data generated from contaminated models

Size	Null			Alternative		
	Normal	t	χ^2	Normal	t	χ^2
Logistic ^a	0.049	0.051	0.052	0.987	0.983	0.985
Huber ^a	0.049	0.048	0.052	0.987	0.983	0.986
Least squares ^a	0.024	0.021	0.019	0.467	0.398	0.404
Logistic ^b	0.054	0.046	0.053	0.884	0.908	0.866
Huber ^b	0.054	0.048	0.052	0.882	0.906	0.862
Least squares ^b	0.021	0.018	0.022	0.455	0.371	0.464

^aThe results are from the case where $\underline{a} \propto \underline{\mu}_2$.

^bThe results are from the case where $\underline{a} \propto (3/2)^{1/2}(1, -1, \dots, 1, -1)^T + (1, \dots, 1, -1, \dots, -1)^T$.

4.3. *Case study.* In this section, we analyze a real microarray dataset and examine the test results based on the least squares method of Feng and He (2009) as well as the robust alternative studied in this paper. We use the same GeneChip data obtained from the MicroArray Quality Control project [Shi et al. (2006), Lin et al. (2013)]. There are a total of 20 microarrays (HG-U133-Plus-2.0) with 54,675 probe-sets (each composed of 11 probes) on each, generated from five colorectal adenocarcinomas and five matched normal colonic tissues with one technical replicate at each of two laboratories involved in the MAQC project. We use the intensity measure of perfect matches, and preprocess the probe-level microarray data with the “RMA” background adjustment method [Irizarry et al. (2003)] and the quantile normalization method [Bolstad et al. (2003)].

We consider a target direction [see supplementary material, Feng and He (2014)] to contrast the two groups: the normal tissue group and the tumor group. Since the gene expressions from the arrays of the same group are expected to be equal, the target direction is approximately orthogonal to the mean of $\underline{\theta}_1$.

For the first approximation by minimizing (2), we use the same values of α and α^* as those of Section 4.2. For the second approximation by minimizing (3), we consider two loss functions: one is for the square loss and the other is the logistic loss function with $C = 1.205$ (times the scale of the residuals). With this choice of C , we retain 95% asymptotic efficiency at the normal distribution.

We inspect one probe-set “1555106_a_at” to better understand the discrepancies between the least squares method and the robust alternative. In this case, the data matrix has 20 rows and $m = 11$ columns. In Figure 1, we

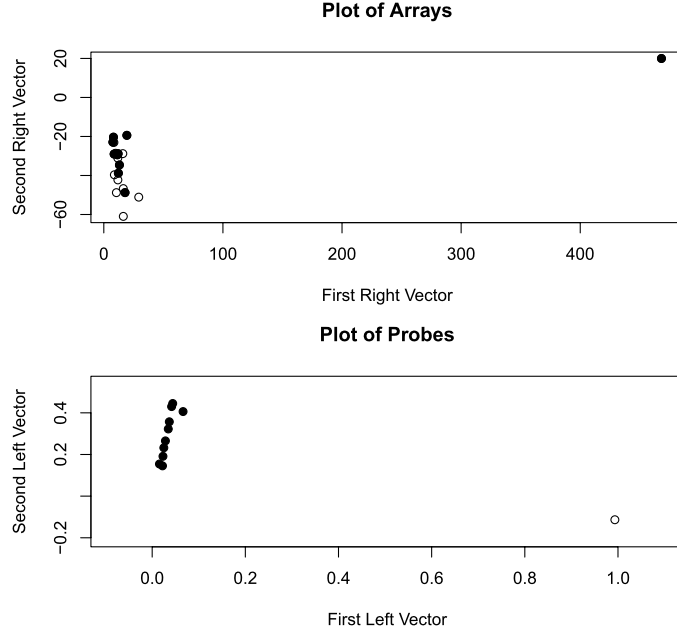


FIG. 1. Scatter plot of singular vectors for the probe-set “1555106_a_at” from the regular SVD. In the upper panel, the circles represent the arrays from tumor samples, while the solid points represent normal tissues. In the lower panel, the circle corresponds to probe 3.

plot the arrays and the probes with the coordinates $(\hat{\theta}_{1i}, \hat{\theta}_{2i})$ and $(\hat{\phi}_{1j}, \hat{\phi}_{2j})$, respectively, for $i = 1, \dots, 20$ and $j = 1, \dots, 11$, where the least squares estimates are used. The p -value is 0.036 based on the least squares method, and the first four singular values are (472, 163, 36, 29). It is clear from Figure 1 that there exist an outlying array and an outlying probe. Further inspection of the data shows that there exists an outlying measurement in the outlying array and the outlying probe in the intensity data matrix. In other words, it is likely that the significant two-dimensional mean structure is caused by the outlier.

With the robust alternative, the p -value is 0.741, and no outlying estimates of the arrays-effects or probe-effects are observed in Figure 2. The first four singular values are (169, 29, 25, 23) in this case, and the second singular value is close to the third and the fourth, which indicates that the 2nd singular structure is likely to be due to noise. From this empirical example, we see that the robust method is powerful in moderating the effect from outliers. More details of the case study can be found in the supplementary material [Feng and He (2014)].

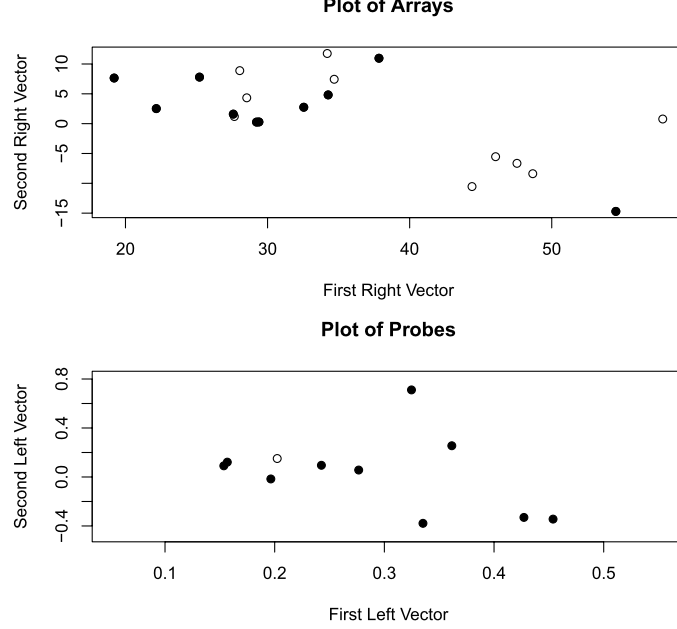


FIG. 2. Scatter plot of singular vectors for the probe-set “1555106_a_at” from a robust approximation. In the upper panel, the circles represent the arrays from tumor samples, while the solid points represent normal tissues. In the lower panel, the circle corresponds to probe 3.

APPENDIX A: ASSUMPTIONS

A.1. Model assumptions.

(M1) The column vectors $\underline{\phi}_k^{(0)}$ ($k = 1, \dots, r$) are orthogonal to one another.

(M2) The row vectors $\underline{\theta}_k^{(0)}$ ($k = 1, \dots, r$) are independently distributed with mean $\underline{\mu}_k = (\mu_{k1}, \dots, \mu_{kn})^T$ and variance $\sigma_k^2 I_n$, for $k = 1, \dots, r$. The components of $\underline{\theta}_k^{(0)}$ are independently distributed with finite fourth moments. Moreover, $\underline{\mu}_k \perp \underline{\mu}_l$, for $k \neq l$, where \perp denotes orthogonality.

(M3) The error variables $\underline{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{im})^T$ are independently generated from a distribution with mean zero and finite fourth moment, and ε_{ij} is symmetrically distributed with $E(\varepsilon_{ij}^2) = \sigma^2$.

(M4) The variables $\{\theta_{1i}^{(0)}\}, \dots, \{\theta_{ri}^{(0)}\}$ and $\{\underline{\varepsilon}_i\}$ are mutually independent.

(M5) $n^{-1} \|\underline{\mu}_k\|^2 \rightarrow \mu_k^2$ as $n \rightarrow \infty$ for some finite constants μ_k , where $\|\cdot\|^2$ is the L_2 norm. We assume that $\mu_k^2 + \sigma_k^2 > \mu_l^2 + \sigma_l^2$ when $k < l$, which is necessary for the identifiability of the model parameters.

These assumptions are clearly satisfied with Gaussian row-effects and Gaussian errors. In robust statistics, a traditional parametric model is often assumed for the outlier-free part of the data, but we design a robust procedure to be insensitive to data contamination.

A.2. Assumptions for Lemma B.2 and Theorem 3.1. Let $\vartheta = (\delta, \varphi^T)^T$, and $\hat{\xi}_\tau$ and ξ_τ be the sample and the population τ quantiles of $\|\hat{\underline{e}}_i\|^2$ and $\|\underline{e}_i\|^2$, respectively, where $\hat{\underline{e}}_i = (I_m - \sum_{k=1}^r \hat{\phi}_k \hat{\phi}_k^T) \underline{y}_i$ and $\underline{e}_i = (I_m - \sum_{k=1}^r \phi_k^{(0)} \phi_k^{(0)T}) \underline{y}_i$. Let the function g denote the probability density of the random variable $\|\underline{e}_i\|^2$.

(E1) The value $g(\xi_\tau)$ is bounded and positive, and g is continuous in a neighbour of ξ_τ .

(E2) $n^{-1} \sum_{i=1}^n E\{\|\underline{y}_i\| \|\frac{\partial f_i}{\partial \underline{y}_i}\| / f_i(\underline{y}_i)\} \leq K$ for some constant K and all n , where f_i is the probability density function of the random vector \underline{y}_i .

(E3) For given $\xi \in \mathbb{R}$, $n^{-1} \sum_{i=1}^n \mathbf{H}_i(\xi, \underline{\phi}_j^{(0)}, \vartheta_0) = O(1)$, for $j = 1, \dots, r$, where

$$(15) \quad \mathbf{H}_i(\xi, \underline{\nu}, \vartheta) = \frac{\partial E\{\mathbf{M}(\xi, \delta, \varphi, \underline{y}_i) \underline{\nu}\}}{\partial \vartheta}$$

and

$$(16) \quad \mathbf{M}(\xi, \delta, \varphi, \underline{y}_i) = 1 \left\{ \left\| \left(I_m - \sum_{k=1}^r \phi_k \phi_k^T \right) \underline{y}_i \right\|^2 \leq \xi + \delta \right\} \underline{y}_i \underline{y}_i^T.$$

REMARK A.1. By similar arguments to those used in the proof of Lemma B.2, we note that assumption (E3) holds if $n^{-1} \sum_{i=1}^n E\{\|\underline{y}_i\|^3 \|\frac{\partial f_i}{\partial \underline{y}_i}\| / f_i(\underline{y}_i)\} \leq K'$ for some constant K' and all n . Conditions (E2) and (E3) are satisfied by the Gaussian distribution as well as any t distribution with finite fourth moment.

A.3. Assumptions on the loss function.

(C1) The loss function L is even and nonnegative, and $L(x) = 0$ if and only if $x = 0$.

(C2) The first derivative L' is continuous, piecewise differentiable, non-decreasing in \mathbb{R} and positive in \mathbb{R}^+ .

(C3) The second derivative L'' is nonnegative, nonincreasing in \mathbb{R}^+ and piecewise continuous.

(C4) The derivatives L' and L'' satisfy $|L'(x)| \leq C_0|x|$ and $L''(x) \leq C_0$ at all $x \in \mathbb{R}$, for some constant C_0 .

A.4. Assumptions for Theorem 4.1.

- (D1) $\max_{1 \leq i \leq n} \|\underline{y}_i\| = O_p(n^{1/4-\delta})$ for some small positive number δ .
 (D2) The distribution of $\theta_{2i}^{(0)} - \mu_{2i}$ is symmetric around zero.

APPENDIX B: PROOFS

In the proofs, we assume that $r = 2$ for simplicity. The same arguments work for the general cases of $r \geq 2$. We first give the Bahadur representations of the quantile estimates. First we state four lemmas, but their proofs can be found in the supplementary material.

LEMMA B.1. *Suppose that assumptions (M1)–(M5) hold and (I2)–(I4) in step 0 are repeated a fixed number of times, then the initial estimate $\hat{\varphi}$ is root- n consistent for φ_0 .*

LEMMA B.2. *Suppose that assumptions (M1)–(M5) and (E1)–(E2) hold, and $\hat{\varphi}$ is the initial root- n consistent estimate of φ_0 , then*

$$(17) \quad \hat{\xi}_\tau - \xi_\tau = -\{ng(\xi_\tau)\}^{-1} \sum_{i=1}^n \psi_\tau\{\|\underline{e}_i\|^2 - \xi_\tau\} + O_p(n^{-1/2}),$$

where $\hat{\xi}_\tau$, ξ_τ , \underline{e}_i , g and v are defined in Appendix A.2, and $\psi_\tau(u) = \tau - 1(u < 0)$.

LEMMA B.3. *If conditions (M2), (M3) and (M5) hold, and $\hat{\varphi}$ is the initial root- n consistent estimate of φ_0 , then*

$$(18) \quad \begin{aligned} & n^{-1} \sum_{i=1}^n 1\{\hat{\xi}_\alpha < \|\hat{\underline{e}}_i\|^2 \leq \hat{\xi}_{1-\alpha}\} \underline{y}_i \underline{y}_i^T \\ & \xrightarrow{p} (1 - 2\alpha)(\mu_1^2 + \sigma_1^2) \underline{\phi}_1^{(0)} \underline{\phi}_1^{(0)T} \\ & \quad + (1 - 2\alpha)(\mu_2^2 + \sigma_2^2) \underline{\phi}_2^{(0)} \underline{\phi}_2^{(0)T} + \sigma^2(\alpha)I, \end{aligned}$$

where $\sigma^2(\alpha) = E[1\{\xi_\alpha \leq \|(I_m - \sum_{k=1}^r \underline{\phi}_k^{(0)} \underline{\phi}_k^{(0)T}) \underline{e}_i\|^2 \leq \xi_{1-\alpha}\} \varepsilon_{ij}^2]$.

Let $\tilde{\varphi}$ be the estimate of φ_0 from step 1. We now have

LEMMA B.4. *Suppose that the observations $\underline{y}_1, \underline{y}_2, \dots, \underline{y}_n$ are drawn from model (4). If assumptions (M1)–(M5) hold, then $\tilde{\varphi} \xrightarrow{p} \varphi_0$.*

In the following lemma, we obtain upper bounds for the estimates of $\underline{\theta}^{(0)}$'s given φ .

LEMMA B.5. *If conditions (C1) and (C2) hold, then we have*

$$(19) \quad f^2(\underline{y}, \underline{\phi}_1, \underline{\phi}_2, \dots, \underline{\phi}_r) + \dots + f^2(\underline{y}, \underline{\phi}_r, \underline{\phi}_1, \dots, \underline{\phi}_{r-1}) \leq 4m^2 \|\underline{y}\|^2$$

for any $\varphi \in \mathbb{S}$ and $\underline{y} \in \mathbb{R}^m$ where

$$(20) \quad \mathbb{S} = \{(\underline{\phi}_1^T, \dots, \underline{\phi}_r^T)^T \in \mathbb{R}^{rm} : \|\underline{\phi}_k\| = 1, \underline{\phi}_k \perp \underline{\phi}_l, \text{ for } k \neq l\}$$

and f is defined in (8)–(9). Furthermore,

$$f^2(\underline{y}, \underline{\phi}_1^*, \underline{\phi}_2^*, \dots, \underline{\phi}_r^*) + \dots + f^2(\underline{y}, \underline{\phi}_r^*, \underline{\phi}_1^*, \dots, \underline{\phi}_{r-1}^*) \leq 4m^3(1 - 3\tau/2)^{-1} \|\underline{y}\|^2,$$

where $\varphi^* = \lambda\varphi_1 + (1 - \lambda)\varphi_2$, $\lambda \in (0, 1)$, and $\|\varphi_1 - \varphi_2\| \leq \tau$ for $0 < \tau < 2/3$ and $\varphi_1, \varphi_2 \in \mathbb{S}$.

REMARK B.1. The result of Lemma B.5 holds uniformly for $\varphi \in \mathbb{S}$, so the existence of the moments of \underline{y}_i ensures the existence of the corresponding moments of the estimates of $\underline{\theta}^{(0)}$'s given φ .

PROOF OF LEMMA B.5. Again we present the proof for $r = 2$. From the definition of f , we have

$$\sum_{j=1}^m L(y_j - f(\underline{y}, \underline{\phi}_1, \underline{\phi}_2)\phi_{1j} - f(\underline{y}, \underline{\phi}_2, \underline{\phi}_1)\phi_{2j}) \leq \sum_{j=1}^m L(y_j),$$

where y_j is the j th component of any vector $\underline{y} \in \mathbb{R}^m$.

From condition (C1), we have

$$L(y_j - f(\underline{y}, \underline{\phi}_1, \underline{\phi}_2)\phi_{1j} - f(\underline{y}, \underline{\phi}_2, \underline{\phi}_1)\phi_{2j}) \leq \sum_{j=1}^m L(y_j) = \sum_{j=1}^m L(|y_j|)$$

for $j = 1, \dots, m$.

We now show that

$$\sum_{j=1}^m L(|y_j|) \leq L\left(\sum_{j=1}^m |y_j|\right).$$

Consider $x, z \in \mathbb{R}$. Without loss of generality, we assume that $x > z > 0$. It is clear that

$$L(x + z) - L(x) = L'(x + \lambda_1 z)z$$

and

$$L(z) - L(0) = L'(\lambda_2 z)z,$$

where $0 < \lambda_1, \lambda_2 < 1$. From conditions (C1) and (C2), we have $L(x+z) - L(x) - L(z) = [L'(x+\lambda_1 z) - L'(\lambda_2 z)]z \geq 0$. Thus, $\sum_{j=1}^m L(|y_j|) \leq L(\sum_{j=1}^m |y_j|)$. It then follows that

$$L(|y_j - f(\underline{y}, \underline{\phi}_1, \underline{\phi}_2)\phi_{1j} - f(\underline{y}, \underline{\phi}_2, \underline{\phi}_1)\phi_{2j}|) \leq L\left(\sum_{l=1}^m |y_l|\right).$$

From (C2), so we have

$$|y_j - f(\underline{y}, \underline{\phi}_1, \underline{\phi}_2)\phi_{1j} - f(\underline{y}, \underline{\phi}_2, \underline{\phi}_1)\phi_{2j}| \leq \sum_{l=1}^m |y_l|.$$

Furthermore, we have

$$|f(\underline{y}, \underline{\phi}_1, \underline{\phi}_2)\phi_{1j} + f(\underline{y}, \underline{\phi}_2, \underline{\phi}_1)\phi_{2j}| \leq \sum_{l=1}^m 2|y_l|.$$

Also note that $\|\underline{\phi}_1\| = \|\underline{\phi}_2\| = 1$ and $\underline{\phi}_1 \perp \underline{\phi}_2$, it then follows that

$$\begin{aligned} & f^2(\underline{y}, \underline{\phi}_1, \underline{\phi}_2) + f^2(\underline{y}, \underline{\phi}_2, \underline{\phi}_1) \\ &= \sum_{j=1}^m [f(\underline{y}, \underline{\phi}_1, \underline{\phi}_2)\phi_{1j} + f(\underline{y}, \underline{\phi}_2, \underline{\phi}_1)\phi_{2j}]^2 \leq 4m \left(\sum_{j=1}^m |y_j| \right)^2 \leq 4m^2 \|\underline{y}\|^2. \end{aligned}$$

With the similar arguments, we have

$$|f(\underline{y}, \underline{\phi}_1^*, \underline{\phi}_2^*)\phi_{1j}^* + f(\underline{y}, \underline{\phi}_2^*, \underline{\phi}_1^*)\phi_{2j}^*| \leq \sum_{j=1}^m 2|y_j|,$$

where $\lambda \in (0, 1)$, $\underline{\phi}_1^* = \lambda \underline{\phi}_1^{(1)} + (1-\lambda)\underline{\phi}_1^{(2)}$ and $\underline{\phi}_2^* = \lambda \underline{\phi}_2^{(1)} + (1-\lambda)\underline{\phi}_2^{(2)}$. Note that

$$\begin{aligned} & \sum_{j=1}^m |f(\underline{y}, \underline{\phi}_1^*, \underline{\phi}_2^*)\phi_{1j}^* + f(\underline{y}, \underline{\phi}_2^*, \underline{\phi}_1^*)\phi_{2j}^*|^2 \\ &= [f^2(\underline{y}, \underline{\phi}_1^*, \underline{\phi}_2^*) + f^2(\underline{y}, \underline{\phi}_2^*, \underline{\phi}_1^*)] \\ &\quad + 2\lambda(1-\lambda)f(\underline{y}, \underline{\phi}_1^*, \underline{\phi}_2^*)f(\underline{y}, \underline{\phi}_2^*, \underline{\phi}_1^*) \\ &\quad \times [\underline{\phi}_1^{(1)T}(\underline{\phi}_2^{(2)} - \underline{\phi}_2^{(1)}) + \underline{\phi}_2^{(1)T}(\underline{\phi}_1^{(2)} - \underline{\phi}_1^{(1)})] \\ &\quad + 2\lambda(1-\lambda)[f^2(\underline{y}, \underline{\phi}_1^*, \underline{\phi}_2^*)\underline{\phi}_1^{(1)T}(\underline{\phi}_1^{(2)} - \underline{\phi}_1^{(1)}) \\ &\quad \quad + f^2(\underline{y}, \underline{\phi}_2^*, \underline{\phi}_1^*)\underline{\phi}_2^{(1)T}(\underline{\phi}_2^{(2)} - \underline{\phi}_2^{(1)})] \\ &\geq (1-3\tau/2)[f^2(\underline{y}, \underline{\phi}_1^*, \underline{\phi}_2^*) + f^2(\underline{y}, \underline{\phi}_2^*, \underline{\phi}_1^*)], \end{aligned}$$

so it follows that

$$\begin{aligned}
& f^2(\underline{y}, \underline{\phi}_1^*, \underline{\phi}_2^*) + f^2(\underline{y}, \underline{\phi}_2^*, \underline{\phi}_1^*) \\
& \leq (1 - 3\tau/2)^{-1} \left(2m \sum_{j=1}^m |y_j| \right)^2 \\
& \leq 4m^3 (1 - 3\tau/2)^{-1} \|\underline{y}\|^2. \quad \square
\end{aligned}$$

LEMMA B.6. *If the result of Lemma B.5 and conditions (C2)–(C4) hold, the following inequality holds for $\underline{\phi}_1$ in a neighbor of $\underline{\phi}_1^{(0)}$,*

$$(21) \quad |L'(y_j - f(\underline{y}, \underline{\phi}_1)\phi_{1j}) - L'(y_j - f(\underline{y}, \underline{\phi}_1^{(0)})\phi_{1j}^{(0)})| \leq C \|\underline{y}\| \|\underline{\phi}_1 - \underline{\phi}_1^{(0)}\|$$

for $j = 1, \dots, m$, where y_j is the j th component of the vector \underline{y} , f is defined in (8)–(9) with $r = 1$, and C is some constant.

PROOF. Without loss of generality, we assume that $\phi_{1j}^{(0)} \neq 0$ for $j = 1, \dots, m_1$, and $\phi_{1j}^{(0)} = 0$ for $j = m_1 + 1, \dots, m$.

(i) Now we consider $j = 1, \dots, m_1$. Consider unit vectors $\underline{\phi}$ and $\underline{\nu}$ such that $\max\{\|\underline{\phi} - \underline{\phi}_1^{(0)}\|, \|\underline{\nu} - \underline{\phi}_1^{(0)}\|\} \leq \tau/2$, where $0 < \tau < 2/3$.

If $L''(y_j - f(\underline{y}, \underline{\phi})\phi_j) = 0$, then

$$\left| \frac{\partial L'(y_j - f(\underline{y}, \underline{\phi})\phi_j)}{\partial \phi_l} \right| = 0.$$

We now consider the case where $L''(y_j - f(\underline{y}, \underline{\phi})\phi_j) > 0$. Let $K_1 = \min\{|\phi_{1j}^{(0)}|, j = 1, \dots, m_1\}$. When $\underline{\phi}$ is sufficiently close to $\underline{\phi}_1^{(0)}$, we must have $|\phi_j| \geq K_1/2$, for $j = 1, \dots, m_1$. It then follows from condition (C3) that

$$\sum_{j=1}^m L''(y_j - f(\underline{y}, \underline{\phi})\phi_j)\phi_j^2 > 0.$$

Note that

$$\sum_{j=1}^m L'(y_j - f(\underline{y}, \underline{\phi})\phi_j)\phi_j = 0,$$

based on the definition of f . By the implicit function theorem, the partial derivatives of f with respect to $\underline{\phi}$ is

$$(22) \quad \frac{\partial f(\underline{y}, \underline{\phi})}{\partial \phi_j} = - \frac{L'(y_j - f(\underline{y}, \underline{\phi})\phi_j) - L''(y_j - f(\underline{y}, \underline{\phi})\phi_j)f(\underline{y}, \underline{\phi})\phi_j}{\sum_{j=1}^m L''(y_j - f(\underline{y}, \underline{\phi})\phi_j)\phi_j^2}$$

for $j = 1, \dots, m$.

Consider the partial derivative

$$\begin{aligned} & \frac{\partial L'(y_j - f(\underline{y}, \underline{\phi})\phi_j)}{\partial \phi_l} \\ &= \begin{cases} -L''(y_j - f(\underline{y}, \underline{\phi})\phi_j)\phi_j \frac{\partial f(\underline{y}, \underline{\phi})}{\partial \phi_l}, & j \neq l, \\ -L''(y_j - f(\underline{y}, \underline{\phi})\phi_j) \left\{ \phi_j \frac{\partial f(\underline{y}, \underline{\phi})}{\partial \phi_l} + f(\underline{y}, \underline{\phi}) \right\}, & j = l. \end{cases} \end{aligned}$$

Let $K_2 = K_1/2$ and

$$z_j(\underline{y}, \underline{\phi}) = \frac{L''(y_j - f(\underline{y}, \underline{\phi})\phi_j)\phi_j}{\sum_{l=1}^m L''(y_l - f(\underline{y}, \underline{\phi})\phi_l)\phi_l^2}.$$

Consider

$$\begin{aligned} & |z_j(\underline{y}, \underline{\phi})| \\ &= K_2^{-1} (L''(y_j - f(\underline{y}, \underline{\phi})\phi_j)|\phi_j|) \\ & \quad \Bigg/ \left(\sum_{j=1}^{m_1} L''(y_j - f(\underline{y}, \underline{\phi})\phi_j)\phi_j^2/K_2 + \sum_{j=m_1+1}^m L''(y_j - f(\underline{y}, \underline{\phi})\phi_j)\phi_j^2/K_2 \right) \\ &\leq K_2^{-1} \frac{L''(y_j - f(\underline{y}, \underline{\phi})\phi_j)|\phi_j|}{\sum_{j=1}^{m_1} L''(y_j - f(\underline{y}, \underline{\phi})\phi_j)|\phi_j|} \leq K_2^{-1}. \end{aligned}$$

It then follows from assumption (C4) and Lemma B.5 that

$$(23) \quad \left| \frac{\partial L'(y_j - f(\underline{y}, \underline{\phi})\phi_j)}{\partial \phi_l} \right| \leq C_1 \{ |f(\underline{y}, \underline{\phi})| + \|\underline{y}\| \} \leq C \|\underline{y}\|$$

for some constant C . Hence, by (C2)–(C4), we obtain

$$(24) \quad |L'(y_j - f(\underline{y}, \underline{\phi})\phi_j) - L'(y_j - f(\underline{y}, \underline{\nu})\nu_j)| \leq C \|\underline{y}\| \|\underline{\phi} - \underline{\nu}\|.$$

(ii) Now consider $j = m_1 + 1, \dots, m$. By condition (C4), we have

$$L''(x) \leq C_0$$

for some constant C_0 , and $x \in \mathbb{R}$. It follows from condition (C3) that

$$\begin{aligned} & |L'(y_j - f(\underline{y}, \underline{\phi})\phi_j) - L'(y_j - f(\underline{y}, \underline{\nu})\nu_j)| \\ &\leq C_0 |f(\underline{y}, \underline{\phi})\phi_j - f(\underline{y}, \underline{\nu})\nu_j| \\ &\leq C_0 \{ |f(\underline{y}, \underline{\phi}) - f(\underline{y}, \underline{\nu})| |\phi_j| + |f(\underline{y}, \underline{\nu})| |\phi_j - \nu_j| \} \\ &\leq C_0 [\{ |f(\underline{y}, \underline{\phi})| + |f(\underline{y}, \underline{\nu})| \} |\phi_j - \phi_j^{(0)}| + |f(\underline{y}, \underline{\nu})| |\phi_j - \nu_j|]. \end{aligned}$$

It then follows from Lemma B.5 that

$$(25) \quad \begin{aligned} & |L'(y_j - f(\underline{y}, \underline{\phi})\phi_j) - L'(y_j - f(\underline{y}, \underline{\nu})\nu_j)| \\ & \leq \|\underline{y}\| (C_2|\phi_j - \phi_j^{(0)}| + C_3|\phi_j - \nu_j|) \end{aligned}$$

for some constants C_2 and C_3 .

Thus, by (24) and (25), we obtain (21). \square

PROOF OF THEOREM 4.1. By (21) and Lemma 4.6 of He and Shao (1996), we have

$$(26) \quad \begin{aligned} & \sup_{|\varphi - \varphi_0| \leq Cn^{-1/2}} \left| \sum_{i=1}^n a_i [\gamma(\underline{y}_i, \varphi) - \gamma(\underline{y}_i, \varphi_0) - E\{\gamma(\underline{y}_i, \varphi) - \gamma(\underline{y}_i, \varphi_0)\}] \right| \\ & = O_p(n^{1/2}), \end{aligned}$$

where $\gamma(\underline{y}_i, \varphi) = \sum_{j=1}^m L'(y_{ij} - f(\underline{y}_i, \underline{\phi}_1)\phi_{1j})\phi_{2j}$.

By the similar arguments to those used to obtain (10) and (11), we obtain $f(\underline{y}_i, \underline{\phi}_1^{(0)}) = \theta_{1i}^{(0)} + f(\theta_{2i}^{(0)}\underline{\phi}_2^{(0)} + \underline{\varepsilon}_i, \underline{\phi}_1^{(0)})$. It then follows from conditions (C1)–(C4), (D2) and (22) that

$$(27) \quad n^{-1} \sum_{i=1}^n a_i \frac{\partial E\{\gamma(\underline{y}_i, \varphi_0)\}}{\partial \varphi_0} = n^{-1} \sum_{i=1}^n a_i E\left\{ \frac{\partial \gamma(\underline{y}_i, \varphi_0)}{\partial \varphi_0} \right\} = o(1),$$

when $\underline{a} \perp \underline{\mu}_1$ and $\underline{\mu}_2 = \underline{0}$. From (23) and (C4), we know that

$$\left| \frac{\partial \gamma(\underline{y}_i, \varphi)}{\partial \varphi} \right| \leq C_1 \|\underline{y}\|$$

for some constants C_1 . It then follows from condition (C3) and the moment condition on \underline{y}_i that $n^{-1} \sum_{i=1}^n a_i \frac{\partial E\{\gamma(\underline{y}_i, \varphi)\}}{\partial \varphi}$ uniformly converges to a continuous function. Thus, it follows from (26) and (27) that

$$(28) \quad \begin{aligned} & \sum_{i=1}^n a_i \left\{ \sum_{j=1}^m L'(y_{ij} - f(\underline{y}_i, \underline{\phi}_1)\phi_{1j})\phi_{2j} \right\} \\ & = \sum_{i=1}^n a_i \left\{ \sum_{j=1}^m L'(y_{ij} - f(\underline{y}_i, \underline{\phi}_1^{(0)})\phi_{1j}^{(0)})\phi_{2j}^{(0)} \right\} + o_p(n^{1/2}). \end{aligned}$$

Under condition (D2) and the null hypothesis that $\underline{\mu}_2 = \underline{0}$, we have

$$n^{-1/2} \sum_{i=1}^n a_i \gamma(\underline{y}_i, \varphi_0)$$

$$= n^{-1/2} \sum_{i=1}^n a_i \left\{ \sum_{j=1}^m L'(\theta_{2i}^{(0)} \phi_{2j}^{(0)} + \varepsilon_{ij} - f(\theta_{2i}^{(0)} \underline{\phi}_2^{(0)} + \underline{\varepsilon}_i, \underline{\phi}_1^{(0)}) \phi_{1j}^{(0)}) \phi_{2j}^{(0)} \right\} \\ \xrightarrow{L} N(0, \alpha^2)$$

as $n \rightarrow \infty$, where

$$\alpha^2 = \text{Var} \left\{ \sum_{j=1}^m L'((\theta_{2i}^{(0)} - \mu_{2i}) \phi_{2j}^{(0)} + \varepsilon_{ij} - f((\theta_{2i}^{(0)} - \mu_{2i}) \underline{\phi}_2^{(0)} + \underline{\varepsilon}_i, \underline{\phi}_1^{(0)}) \phi_{1j}^{(0)}) \phi_{2j}^{(0)} \right\}.$$

Note that

$$|\gamma^2(\underline{y}_i, \tilde{\varphi}) - \gamma^2(\underline{y}_i, \varphi_0)| \leq K \|\underline{y}_i\| \|\tilde{\varphi} - \varphi_0\|$$

for some constant K , so $\tilde{\sigma}_n^2 \xrightarrow{p} \alpha^2$ as $n \rightarrow \infty$, under the null that $\underline{\mu}_2 = \underline{0}$. Therefore, we obtain (12). \square

Acknowledgements. We thank anonymous reviewers, including an Associate Editor, who provided valuable critiques about earlier versions of this paper. Their comments motivated us to find better robust low-rank approximations to data matrices with a solid theoretical underpinning. They also helped us improve the presentation.

SUPPLEMENTARY MATERIAL

Additional details of case study and technical proofs

(DOI: [10.1214/13-AOS1186SUPP](https://doi.org/10.1214/13-AOS1186SUPP); .pdf). We provide details of the case study in Section 4.3 and complete the proofs of technical lemmas, as well as Theorems 3.1–3.2 and 4.2 of this paper.

REFERENCES

- AGARWAL, A., NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Ann. Statist.* **40** 1171–1197. [MR2985947](#)
- AMMANN, L. P. (1993). Robust singular value decompositions: A new approach to projection pursuit. *J. Amer. Statist. Assoc.* **88** 505–514. [MR1224375](#)
- BOLSTAD, B. M., IRIZARRY, R. A., ASTRAND, M. and SPEED, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19** 185–193.
- CANDÈS, E. J., LI, X., MA, Y. and WRIGHT, J. (2011). Robust principal component analysis? *J. ACM* **58** Art. 11, 37. [MR2811000](#)
- CHEN, C., HE, X. and WEI, Y. (2008). Lower rank approximation of matrices based on fast and robust alternating regression. *J. Comput. Graph. Statist.* **17** 186–200. [MR2424801](#)
- FENG, X. and HE, X. (2009). Inference on low-rank data matrices with applications to microarray data. *Ann. Appl. Stat.* **3** 1634–1654. [MR2752151](#)

- FENG, X. and HE, X. (2014). Supplement to “Statistical inference based on robust low-rank data matrix approximation.” DOI:[10.1214/13-AOS1186SUPP](https://doi.org/10.1214/13-AOS1186SUPP).
- GABRIEL, K. R. and ZAMIR, S. (1979). Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics* **21** 489–498.
- GERVINI, D. and YOHAI, V. J. (2002). A class of robust and fully efficient regression estimators. *Ann. Statist.* **30** 583–616. [MR1902900](#)
- HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, 1st ed. Wiley, New York. [MR0829458](#)
- HE, X. and SHAO, Q.-M. (1996). A general Bahadur representation of M -estimators and its application to linear regression with nonstochastic designs. *Ann. Statist.* **24** 2608–2630. [MR1425971](#)
- HUBER, P. J. and RONCHETTI, E. M. (2009). *Robust Statistics*, 2nd ed. Wiley, Hoboken, NJ. [MR2488795](#)
- IRIZARRY, R., BOLSTAD, B. M., COLLIN, F., COPE, L. M., HOBBS, B. and SPEED, T. P. (2003). A model-based background adjustment for oligonucleotide expression arrays. *Nucleic Acids Res.* **31** e15.
- LI, C. and WONG, W. H. (2001). Model-based analysis of oligonucleotide arrays: Expression index and outlier detection. *Proc. Natl. Acad. Sci. USA* **98** 31–36.
- LIN, G., HE, X., JI, H., SHI, L., DAVIS, R. W. and ZHONG, S. (2013). Reproducibility probability score—incorporating measurement variability across laboratories for gene selection. *Nat. Biotechnol.* **24** 1476–1477.
- MAMMEN, E. (1991). *When Does Bootstrap Work? Asymptotic Results and Simulations*, 1st ed. Springer, New York.
- NANDAKUMAR, R., YU, F., LI, H. and STOUT, W. (1998). Assessing unidimensionality of polytomous data. *Appl. Psychol. Meas.* **22** 99–115.
- RUPPERT, D. and CARROLL, R. J. (1980). Trimmed least squares estimation in the linear model. *J. Amer. Statist. Assoc.* **75** 828–838. [MR0600964](#)
- SHI, L., REID, L. H., JONES, W. D. et al. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* **24** 1151–1161.
- ZHOU, Z., LI, X., WRIGHT, J., CANDES, E. and MA, Y. (2009). Stable principal component pursuit. In *International Symposium on Information Theory*, June 2010.

SCHOOL OF STATISTICS AND MANAGEMENT
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS
AND KEY LABORATORY OF MATHEMATICAL ECONOMICS (SUFE)
MINISTRY OF EDUCATION
777 GUODING ROAD
SHANGHAI 200433
CHINA
E-MAIL: feng.xingdong@mail.shufe.edu.cn

DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
439 WEST HALL
1085 SOUTH UNIVERSITY AVENUE
ANN ARBOR, MICHIGAN 48109
USA
E-MAIL: xmhe@umich.edu